ELSEVIER

Computers and Chemistry 24 (2000) 489–497

Computers & Chemistry

www.elsevier.com/locate/compchem

A global optimization strategy for predicting α-helical protein tertiary structure[☆]

Silvia Crivelli [a,c], Richard Byrd [d], Elizabeth Eskow [d], Robert Schnabe [d], Richard Yu [b,1], Thomas M. Philip [b], Teresa Head-Gordon [a,b,*]

[a] *Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, 94720 California, USA*
[b] *Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, 94720 California, USA*
[c] *NERSC Division, Lawrence Berkeley National Laboratory, Berkeley, 94720 California, USA*
[d] *Department of Computer Science, University of Colorado, Boulder, CO, USA*

Received 16 August 1999; received in revised form 4 October 1999; accepted 4 October 1999

Abstract

We present a global optimization strategy that incorporates predicted restraints in both a local optimization context and as directives for global optimization approaches, to predict protein tertiary structure for α-helical proteins. Specifically, neural networks are used to predict the secondary structure of a protein, restraints are defined as manifestations of the network with a predicted secondary structure and the secondary structure is formed using local minimizations on a protein energy surface, in the presence of the restraints. Those residues predicted to be coil, by the network, define a conformational sub-space that is subject to optimization using a global approach known as stochastic perturbation that has been found to be effective for Lennard–Jones clusters and homo-polypeptides. Our energy surface is an all-atom 'gas phase' molecular mechanics force field, that is combined with a new solvation energy function that penalizes hydrophobic group exposure. This energy function gives the crystal structure of four different α-helical proteins as the lowest energy structure relative to other conformations, with correct secondary structure but incorrect tertiary structure. We demonstrate this global optimization strategy by determining the tertiary structure of the A-chain of the α-helical protein, uteroglobin and of a four-helix bundle, DNA binding protein. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Conformational search; Constraint optimization; Parallelization; Hydrophobic effect; Energy functions

1. Introduction

The protein structure prediction problem is to determine the three-dimensional arrangement of the protein molecule, given a protein-solvent potential or free energy surface in accordance with the amino acid sequence (Vasquez et al., 1994; Eisenhaber et al., 1995). The 'rugged landscape' topography of this surface, defines the underlying difficulty in solving the protein structure prediction problem; the native structure minimum, presumably the global minimum, must be discriminated from other minima, whose number rises

[☆] Workshop on New Trends in Computational Methods for Large Molecular Systems, Szklarska Poreba, Poland, 1–6 July 1999, Edited by W. Andrzej Sokalski and Morris Krauss.

[*] Corresponding author. Tel.: + 1-510-486-7365; fax: + 1-510-486-6488.

E-mail address: tlhead-gordon@lbl.gov (T. Head-Gordon)

[1] Present address. Department of Biophysics, Yale University, New Haven, CT, USA.

exponentially with the number of amino acids in the sequence. Furthermore, this energy surface is difficult to model reliably in a global sense, i.e. to ensure that all misfolds are higher in energy than the correctly folded conformation.

For a sufficiently well-defined energy surface, mathematical optimization research for obtaining the global solution to a large nonlinear system with numerous local minima, can be broadly categorized into two approaches. Constrained optimization methods rely on the availability of sufficiently well-defined constraints, so that the desired solution is the only available solution and local optimization algorithms can be applied. However, the necessary set of biophysical constraints, needed for robust protein structure determination cannot be unambiguously predicted at this time (Head-Gordon et al., 1991, 1992; Gay et al., 1992; Head-Gordon and Stillinger, 1993a,b; Head-Gordon, 1994). In fact, various conformational search strategies assume perfect knowledge of some aspects of the structure: for example α-helical and β-sheet secondary structures (Friesner and Gunn, 1996). Global optimization techniques, in principle, avoid this predictive capacity problem, by systematically searching the potential energy surface to find all low-lying minima including the global energy minimum. Global optimization approaches are much more immature, by comparison to local minimization and theoretical guarantees for finding the global minimum are non existent or weak guarantees at best.

Global optimization methods only make sense in the context of an objective function whose global minimum is actually the desired minimum sought. While in principle the protein structure prediction problem seeks the global free energy minimum, because folding is under thermodynamic control (although there are many examples of proteins that are long-lived stable kinetic intermediates), the modeling of an energy function. That guarantees the native structure as the global minimum relative to all conceivable misfolded structures is a formidable task. It is well appreciated that so-called ‘gas phase’ protein molecular mechanics force fields do not differentiate well between folded and misfolded energy conformations (Novotny et al., 1984). Qualitative improvement in lowering the energy of correctly folded structures relative to non native structures, is to incorporate a description of aqueous solvation (Wesson and Eisenberg, 1992; Schiffer et al., 1993).

We introduce a global optimization strategy and a new energy function that describes the hydrophobic effect, to predict the structure of α-helical proteins. Our global optimization approach is to make good predictions of certain aspects of protein structure such as α-helices, β-sheets and coil regions by neural network techniques (Qian and Sejnowski, 1988; McGregor et al., 1989; Kneller et al., 1990; Holley and Karplus, 1991; Muskal and Kim, 1992; Head-Gordon and Stillinger, 1993a,b; Stillinger et al., 1993; Rost and Sander, 1994; Yu and Head-Gordon, 1995) and then manifest them as restraints to use within both a local optimization algorithm and as guidance within various global optimization frameworks (Byrd et al., 1994, 1995a,b; Azmi et al., 1999). The use of restraints should allow the local minimization components of the method to quickly refine α-helices and β-sheets, when they are predicted with reasonable accuracy. The global optimization component is a stochastic-perturbation algorithm that minimizes in dihedral angles (Byrd et al., 1994, 1995a,b), with a key new component being steps that perform global optimizations over small subsets of dihedral angles that are predicted to be coil from the network predictions (Azmi et al., 1999; Crivelli and Head-Gordon, 1999; Crivelli et al., 1999). Global optimization should be particularly effective in resolving these regions for which it is not possible to define a soft constraint.

In addition, we introduce an atomic pairwise additive solvation term, that stabilizes the burial of hydrophobic groups as well as spatially longer ranged stabilization of hydrophobic groups when there are interleaving polar atoms or water. When this solvation potential is combined with the AMBER95 protein force field (Cornell et al., 1995), the total function gives the crystal structure of four different α-helical proteins as the lowest energy structure, relative to other conformations with correct secondary structure but incorrect tertiary structure.

This paper is meant to describe our methodology and developed algorithm for ab initio prediction of protein tertiary structure with combined restraints and stochastic-perturbation with a reliable energy function. We have recently reported on preliminary results found using an analytical smoothing technique, combined with stochastic-perturbation and restraints (Azmi et al., 1999). It is important to emphasize that this work is very preliminary, but robust enough to provide a specific ab initio prediction on two α-helical proteins, the A chain of uteroglobin (2utg_A) (Bally and Delettre, 1984) and a four helix bundle DNA binding protein (1pou) (Assa-Munt et al., 1993). The promise of our approach will only be known with further testing and predictions of a large number of α-helical proteins and extensions to β-sheet topologies.

2. Methods

2.1. The energy function

The AMBER molecular mechanics energy function (Cornell et al., 1995), V_{MM}, is used to represent the protein–protein interactions. We have also added an empirical solvation free energy term, $V_{\text{solvation}}$, to describe hydrophobic effects that acts between all aliphatic carbon centers. This description is motivated by our recent experimental, theory and simulation work to determine the role of hydration forces in the folding of model protein systems (Pertsemlidis et al., 1996; Head-Gordon et al., 1997; Sorenson and Head-Gordon, 1998; Hura et al., 1999; Pertsemlidis et al., 1999; Sorenson et al., 1999), as well as Pratt–Chandler integral equation theory (Pratt and Chandler, 1977, 1980) that describes solute–solute correlations for small hydrophobic solutes in aqueous solution. The integral equation theories and simulations (Zichi and Rossky, 1985; Head-Gordon, 1995; Rick and Berne, 1997) for the association of two methane molecules in water, show that there are two free energy minima for the molecules in contact and the molecules separated by a length-scale of one water molecule, with a barrier in between. The benefit of this description is that (1) we introduce a stabilizing force for forming hydrophobic cores; (2) it is a well-defined model of the hydrophobic effect for hydrophobic groups in water; and (3) it can be devised as a continuous potential that is computationally tractable relative to solvent accessible surface area models (Schiffer et al., 1993).

The functional form of the solvation term is a sum of gaussians

$$V_{\text{solvation}} = \sum_{i}^{N_C}\sum_{j}^{N_C}\sum_{k}^{M} h_k \exp\left(-\left(\frac{r_{ij}-c_k}{w_k}\right)^2\right) \tag{1}$$

where the sum over i and j is over the aliphatic carbon centers and the sum over k is the number of gaussians necessary to describe the position (c_k), depth (h_k) and width (w_k) of the minima and barrier of the aqueous methane potential of mean force. For uteroglobin, we first tested a solvation energy function, using values of c_k, h_k, w_k and M that reproduced the potential of mean force of two methanes in water taken from a molecular simulation using a novel representations of liquid water (Head-Gordon, 1995). However, we initially found better agreement between good folds and energies by eliminating the solvent-separated minimum, i.e. we have used values of the parameters for stabilizing methanes in contact only.

Table 1
Parameters used for solvation function, $V_{\text{solvation}}$ Eq. (1)

Gaussian, i	C_i	H_i	W_i
1	3.33343	0.73692	1.59315
2	4.94296	0.30938	0.71909
3	6.65118	−0.57310	0.63525
4	1.83343	−2.23692	0.59315

For the DNA binding protein, our new optimized solvation energy function used a potential of mean force description with the solvent-separated minimum restored, but more exaggerated stabilization of both contact and solvent separated minimum with respect to the original methane potential of mean force. Electrostatic interactions were also screened by a dielectric constant of 4, typical of a protein environment. Furthermore, this energy function gives the crystal structure of four different α-helical proteins (including uteroglobin) as the lowest energy structure relative to other conformations with correct secondary structure but incorrect tertiary structure that we have found so far. The proteins are 2utg_A (Bally and Delettre, 1984), 1pou (Assa-Munt et al., 1993), 3icb (Szebenyi and Moffat, 1986) and calmodulin and we have interrogated on the order of 40 000 structures (with a majority for 1pou and 2utg_A) that were all higher in energy than the crystal structure, with root mean square deviations (r.m.s.d.) between α-carbons ranging from 5.5 to 15.0 Å. Table 1 contains the parameters and functional form used for solvation in our present study.

2.2. Neural network algorithm

Our research in neural network prediction of protein secondary structure, has focused on the design of neural network architectures, that actually mitigate the degradation of network performance, due to database deficiencies and the multiple minimum problem in the space of the network variables. Thus far, we have considered network architecture design for helix/no helix prediction of real proteins (Head-Gordon and Stillinger, 1993a,b), a pilot study of secondary structure prediction for real proteins using an input window of nine amino acids (Yu and Head-Gordon, 1995) and tertiary structure for complete sequence-structure databases of a model chemistry (Head-Gordon and Stillinger, 1993a,b). In the pilot study of protein secondary structure we showed that, compared to arbitrary network architectures, network design features serve as constraints for a more optimal network solution, that partially overcomes the network multiple minima problem (Yu and Head-Gordon, 1995). These designed networks also exhibited superior generalization to the test set of proteins, partially overcoming the deficiencies of the training database, by more efficiently mining general rules and not specifics of the training set of proteins. Below, we outline our neural network topology approach for secondary structure prediction, for real protein databases, using an input window size of 17 amino acids, with a two-bit, three-state output (helix, sheet and coil). We hope to provide an expanded report on this work in the near future.

The network involves a careful choice of input representation for each amino acid, a primary structure input window with individual feature detectors of local secondary structure, as well as hidden neurons that amplify whether the window is composed of helix-promoting or sheet-promoting residues. The network also has a recurrent input space comprising the current secondary structure prediction for each amino acid in the sequence window. Finally, we assume imperfect knowledge, as to whether a protein is classified as all-helix, all-sheet, or other. Our results for our primary network can be summarized as follows: when the threshold for the two-state output maximizes the percentage correct on the training set, we obtain an average prediction on nine test sets of $Q_{\text{tot}} = 66.35\%$, with $Q_{\text{a}} = 58.17$, $C_{\text{a}} = 0.48$, $Q_{\text{b}} = 45.85$, $C_{\text{b}} = 0.40$ and $Q_{\text{coil}} = 77.52$, $C_{\text{coil}} = 0.40\%$, where C_{a} and C_{b} are correlation coefficients, Q_{a}, Q_{b} and Q_{coil} are percentages of correct predictions of α-helices, β-sheets and coils, respectively and Q_{tot} is the sum of the percentages above defined.

It is important to emphasize that we have achieved this performance (1) without using sequence or structural homologies as either input or training parameter; (2) without a 'jury' of networks, or special criteria for selecting the best trained network; and (3) using far fewer network variables than past feed-forward back-propagation networks, that predict secondary structure. We emphasize these points, not to diminish the importance of past efforts, but to indicate that the combination of our fully designed networks with sequence homology and trained network selection, may actually boost average performance above the current average of $\approx 70\%$ (Rost and Sander, 1994) in the future.

We have focused on ten α-helical target proteins that range in size between ≈ 70 and 150 amino acids for prediction, using our global optimization strategy, although we only have results on two proteins thus far: 2utg_A and 1pou. Our overall secondary structure prediction performance on these ten proteins, range from 67 to 88% correct secondary structure assignment. For 2utg_A and 1pou the secondary structure predictions are 80% correct.

2.3. The use of 'soft' constraints in protein structure prediction

We have applied predicted structural information in energy minimization predictions in the 'antlion' method (Head-Gordon et al., 1991; Head-Gordon and Stillinger, 1993a,b; Head-Gordon, 1994). The ultimate objective of this method, is to simplify the energy surface for any polypeptide or protein, so that only a single minimum remains. Furthermore, the remaining minimum should occur 'close' to the initial hypersurface native structure minimum. Optimization then proceeds in three stages: (1) use predicted structure information to replace the complicated hypersurface by its simplified variant; (2) optimize on the simplified hypersurface; (3) optimize on the 'real' hypersurface using the optimized structure found from the second stage as an initial guess.

The operation for smoothing the energy surface is to formulate mathematical functions that are added on as biases or restraints to the original surface. These functions are derived from imperfect and incomplete protein structure prediction, based on other methods such as neural networks. The protein structure biasing method emphasizes that a local optimization algorithm with well-formulated predicted constraints or biases can successfully deal with some aspects of the global optimization problem. The biasing method has been demonstrated to be successful on a small, naturally occurring 26-residue polypeptide, melittin, which forms two α-helices separated by a bend at mid-sequence (Head-Gordon and Stillinger, 1993a,b; Head-Gordon, 1994).

Given the neural network predictions of the secondary structure state of each amino acid, for a given protein, two restraints can be defined for α-helical and β-sheet categories. The first is a bias of the backbone torsional angles of a residue according to

$$V_{\phi\psi} = k_{\phi}(1 - \cos(\phi - \phi_0)) + k_{\psi}(1 - \cos(\psi - \psi_0)) \qquad (2)$$

where ϕ_0 and ψ_0 are assigned values appropriate to a perfect α-helix or β-sheet and k_{ϕ} and k_{ψ} are force constants related to the output, or strength, of the neural network. For the case of helical proteins, the focus of this paper, amino acids that are predicted to be non helical will have small force constants, while residues predicted to be helical will have force constants that give rise to stronger restraints. The second function

$$V_{\text{HB}} = q_i q_{i+4} / r_{i,i+4} \qquad (3)$$

encourages helical hydrogen bonds to form between the oxygen atom of residue i, O_i, and the hydrogen of residue $i+4$, H_{i+4}. In this case $q_i = -q_{i+4}$, is the direct neural network output and provides a strong incentive for an intramolecular hydrogen bond to form when residue i is strongly predicted to be helical. Development of restraints such as Eq. (2) and Eq. (3) for β-sheets is to be explored by 'matching' algorithms (Lovasz and Plummer, 1986) in the near future.

2.4. Stochastic-perturbation global optimization algorithm

Given the predictions of secondary structure described above, there remains the difficult optimization problem of finding the torsion angles not specified by those predictions, as well as determining the values of the predicted angles more precisely. In this problem, the potential energy function still has a very large number of local minimizers and a good large-scale global opti-

Table 2
Outline of stochastic-perturbation algorithm

Phase I: generation of initial configurations

(1) Generate sample configuration.
Build up sample configurations by sequentially generating random values for each dihedral angle (from one end of the protein to the other) and choosing the one that produces the lowest energy function.

(2) Optimize.
Select a subset of the best configurations created in step (1): perform a full-dimensional local minimization for each configuration created: store a subset of the best minimizers for further improvement in Phase II.

Phase II: global optimization.

For some number of iterations:

(1) Define a small-dimensional subproblem.
Select a minimizer from the list of full-dimensional local minimizers and a small subset of dihedral angles to be optimized.

(2) Perform a small-scale global optimization
To find the best values for the dihedral angles selected in Step 1 with the remaining angles temporarily fixed at their current values.

(3) Refine the best structure resulting from Step 2
Into local minimizers in the full variable space, using roughly the same process as in Step 2 of Phase I and merge the new lowest configurations into the existing list of local minimizers.

mization algorithm is required. We carry out this optimization using the stochastic-perturbation algorithm developed in (Byrd et al., 1994, 1995a,b) which is based on the original approach of Rinnooy-Kan (Rinnooy-Kan and Timmer, 1984). The two novel aspects of this method are that redundant work is avoided, by assigning new sample points to basins of attraction, defined within a critical radius that avoids minimizations that reach the same local minimum (Rinnooy-Kan and Timmer, 1984; Byrd et al., 1994, 1995a,b). Furthermore, the method has theoretical guarantees of finding global minimum when enough sample points are used, although high probabilities of finding global minimum are prohibitively expensive.

The stochastic-perturbation global optimization algorithm, is based on generating and improving a pool of local minimizers of the objective function. It consists of two phases. In the first phase (phase I), a set of initial conformations is randomly generated and each is used as a starting point for a local minimization. The best of the resulting local minimizers forms a pool used in the next phase. The second phase (phase II) consists of repeatedly selecting the best unexamined conformation and modifying it, using a small-dimensional global optimization program. This global optimization is done holding all internal coordinates of the chosen conformation fixed except for 2–10 of the torsion angles. We use an adaptation of the probabilistic algorithm (Rinnooy-Kan and Timmer, 1984) for this small-dimensional optimization. About 5–25 of the best local minimizers, found in the small-dimensional global optimization, are then used as starting points for local optimization over all problem variables using a limited memory quasi-Newton method (Liu and Nocedal, 1989). The phase II iteration is repeated, if resources permit, until no further progress can be made. A framework for the stochastic-perturbation algorithm is outlined below in Table 2.

The stochastic-perturbation algorithm allows one to explore the vast search space of possible configurations alternatively in breadth and depth. The configurations passed from phase I to phase II can be thought of as the roots of trees of possible solutions, that are first traversed in depth regardless of the energy function values, when compared across the breadth of the tree. This is important, as the energy values do not necessarily decrease monotonically as the tree is traversed in depth. After the initial phase, in which all the trees have been searched to some specified depth, the selection of configurations for the second phase is based on the energy value. The number of these configurations to be considered for further refinement will determine the breadth of the search.

2.5. Combining restraints and the stochastic-perturbation approach

The novel contribution of this research, is the use of partial secondary structure information within a global optimization algorithm, for determining tertiary structure. As the starting point for this approach, secondary structure is predicted by the neural network algorithm described above. Following neural network prediction of secondary structure, our approach, like the previous stochastic-perturbation algorithm, consists of two phases.

The first phase starts with a completely extended conformer, with no secondary or tertiary structure and performs local minimizations using the sum of V_{MM}, $V_{\text{solvation}}$ and the restraints defined in Eq. (2) and Eq. (3) first and then the unbiased potential energy function, $V_{MM} + V_{\text{solvation}}$. The local minimizations on the biased function encourage the formation of a-helices in the regions where predictions of α-helix are strong. It is important to mention that, because the network predictions may not be exact, the biasing terms may either force some helical forms in places where they do not belong or discourage their formation in places where they do belong. The local minimizations on the unconstrained function allow the entire configuration to change, in an attempt to correct at least partially, those areas in which predictions are wrong. The typical output from this phase is at least partially correct in its secondary structure, but does not contain correct tertiary structure.

The second phase starts with the outcome of the previous phase as the first member of a list of local minimizers. From the set of dihedral angles predicted to be coil, the algorithm randomly selects a subset. The algorithm then performs a small-scale global optimization, using the selected dihedral angles as variables, while keeping the rest temporarily fixed at their current values. This optimization produces a number of local minimizers to the unbiased energy function, in the subspace of dihedral angles chosen and then through assignment of minima to basins of attraction and selected minimizations, returns the global minimum in that sub-space. The algorithm has been parallelized (Crivelli and Head-Gordon, 1999; Crivelli et al., 1999) so that on the order of 5–10 subspaces can be explored at once. A number of those conformations with low energy values are considered for further refinement, that is done by performing local minimizations on the full variable space. These minimizations are performed using the unconstrained energy function. The new minimizers obtained from the local minimizations are merged into the current list of minimizers. The lowest energy conformation is selected from this list and the second phase starts again. The process repeats for a number of iterations. The new stochastic-perturbation algorithm is outlined in Table 3.

Table 3
Outline of stochastic-perturbation with biasing algorithm

Phase I: generation of initial configurations based upon structure prediction

(1) Generate sample configuration.
Start with completely extended conformer: no secondary or tertiary structure.
(2) Optimize with restraints
(a) Create the helical biasing terms, Eq. (2) and Eq. (3): force constants and charges are defined as the output of the neural network (a value between 0.0 and 1.0); force constants are scaled to appropriate energy units.
(b) Perform local minimization on the biased energy function which incorporates structure prediction in the output of (2a).
(c) Perform local minimization on the unbiased potential energy function using the structures from (2b) as starting configuration.

Phase II. global optimization utilizing structure prediction.

For some number of iterations:
(1) Define a small scale global optimization problem based on predictions.
Select a configuration from the list of local minimizers: select a subspace, defined as 4–10 dihedral angles randomly chosen from the amino acids predicted to be coil.
(2) Perform a small-scale global optimization
To find the best values for the dihedral angles selected in step 1b with the remaining angles temporarily fixed at their current values. This stage uses the unbiased energy function.
(3) Refine the few best structures resulting from step 2
Into local minimizers in the full variable space that are consistent with the structure predictions and merge the new lowest configurations into the existing list of local minimizers.

3. Results

We test the stochastic-perturbation with restraints algorithm on the prediction of the A-chain of uteroglobin, 2utg_A and a four-helix bundle DNA binding protein, 1pou, proteins comprised of ≈ 70 amino acids. Phase I is primarily the generation of a structure (or structures) consistent with the restraints. It begins with a starting structure of the target protein, that is the minimum closest to the fully extended form with all backbone pairs assuming $\phi = 180°$ and $\psi = -180°$. The optimization is performed on the sum of the AMBER surface, V_{MM} and $V_{\text{solvation}}$ and the restraints from the neural network predictions (Eq. (2) and Eq. (3)), using a limited memory BFGS local minimization algorithm (Liu and Nocedal, 1989). The converged structure on the modified surface is then used as a starting configuration for optimization on $V_{MM} + V_{\text{solvation}}$ alone. The number of iterations necessary to meet this goal is about 5000.

Fig. 1 shows a ribbon diagram comparison between the crystal structure of the structural target 2utg_A and the end product of phase I of our algorithm. At the end of phase I, is a protein with its helices formed and extended structures in regions predicted to be coil. The neural network made good predictions of the helical content of the target protein; three of the four helices in this protein are reasonably well formed as a result. One helix is much more distorted, however, as a consequence of weak predictions in that region of the sequence. It is important that the network did predict these regions to be helical, but only weakly so. We did another test with a modified prediction file that assumes stronger helix prediction, i.e. by increasing the scaling factor in that region of the sequence and obtained a better formed helix. We decided to use the latter structure as a starting configuration for phase II, as well as the modified prediction file, as we anticipate that future improvements in our network prediction algorithm,

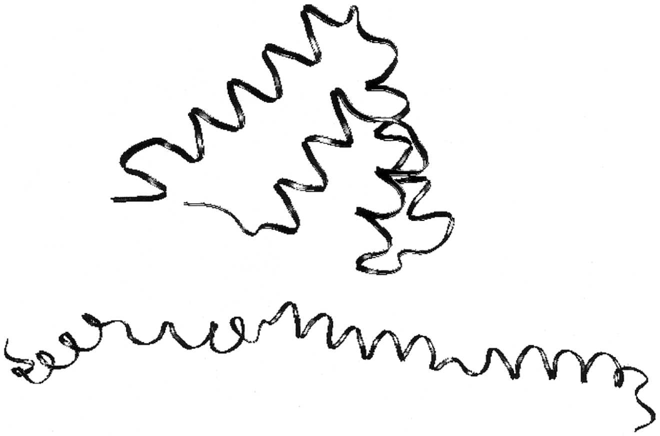

Fig. 1. A ribbon diagram comparison between the crystal structure of the A-chain of uteroglobin (top) and the end product of phase I of our algorithm (bottom). At the end of phase I is a conformation with its helices formed and extended structures in regions predicted to be coil. The neural network made good predictions of three of the four helices in this protein and are reasonably well formed as a result. One helix was much more distorted, however, as a consequence of weak predictions in that region of the sequence. We modified four helical predictions in this helix to be stronger and used the modified predictions to arrive at the structure shown here.

largely exploiting multiple alignments, will largely correct weak predictions for α-helical proteins. We note that at the end of phase I for this algorithm, essentially no structural diversity is introduced into the start of phase II. Similar results were found for 1pou, although we did not need to modify the predictions in any way.

It is possible to either use the result of phase I directly as the starting structure of phase II, or generate random values for the backbone dihedral angles of the predicted coiled regions to incorporate structural diversity at the end of phase I. We used the result of phase I as the starting configuration for phase II for 2utg_A and we generated a diversity of starting structures from the phase I output for 1pou by randomizing coil dihedral angle values.

An iteration of phase II of the algorithm performs a global optimization in a sub-space of some dihedral angles that are chosen from all residue ϕ, ψ, χ torsion triplets (2utg_A) and ϕ, ψ torsion pairs (1pou) predicted by the network to be coil. Within this subspace, the global optimization method of Rinnooy-Kan, systematically explores this space to zone in on the region most likely to contain the global minimum. The algorithm is general, in the sense that arbitrary dimensional sub-space sizes can be explored. We have decided on a strategy of defining the set of 28 predicted coil residues as the pool of possible ϕ, ψ pairs or ϕ, ψ, χ triplets and either three pairs (1pou) or two triplets (2utg_A) were randomly chosen from that pool, for a total sub-space of six dihedral angles.

The chosen sub-space is subdivided in M regions, M being the number of workers. Each worker randomly generates 50 sample points over a uniform distribution, in its assigned region of the domain, for a total of 400 points on the entire subspace (Crivelli and Head-Gordon, 1999; Crivelli et al., 1999). Each phase II iteration corresponds to roughly 4 cpu hours using between 28 and 64 processors on the Cray T3E.

Fig. 2a shows the best energy conformer for 2utg_A after ten iterations of phase II, where the overall shape of the predicted fold is quite reasonable. The α-carbon r.m.s.d., between the backbone of the crystal structure and the backbone of the predicted structure is just under 7.4 Å. Since the algorithm is statistical, i.e. perturbations of the sub-space definitions are random, we may need more runs to verify that no other structure is found. However, part of the output of a phase II

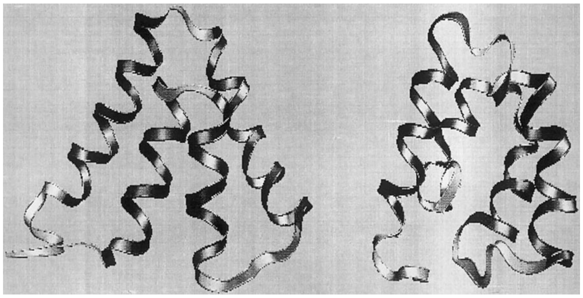

Fig. 2. (a) A ribbon diagram comparison between the crystal structure of the A-chain of uteroglobin (left, red) and the outcome of phase II of our algorithm (right, blue). (b) A ribbon diagram comparison between the NMR structure of a four helix bundle DNA binding protein (1pou) (right, red) and the outcome of phase II of our algorithm (left, blue).

run is a list of the 20 most energetically favorable structures found. Some of these structures differ in small ways from the best energy conformer and are only slightly higher in energy on average. In cases where the energies are very different, the fold is very different as well, indicating that the energy function is acting sensibly.

Fig. 2b shows the best prediction for 1pou. The α-carbon r.m.s.d. value of our prediction with respect to the NMR experimental structure is 6.3 Å. An important benefit of our new energy function is that our current best prediction is higher in energy than the NMR structure, which means that further optimization may lower the α-carbon r.m.s.d. even further as energy improves. Finally, we note that when an all heavy atom r.m.s.d. is evaluated, i.e. including side chains, that the r.m.s.d's, increase only 0.5–1.0 Å, indicating that roughly, correct tertiary fold found by conformational searches over the backbone degrees of freedom influences reasonable side chain packing.

4. Summary and conclusions

We have developed and tested our new methodology described here for determining tertiary structure of α-helical proteins. Neural network predictions of secondary structure are manifested as restraints that permit partial solution to the global optimization problem within a local optimization algorithm. The neural networks also bridge the gap between primary and tertiary structure, by greatly narrowing the conformational search space, by focusing the work in the subspace of dihedral angles predicted to be coil by the neural network.

We have been able to find, based on the AMBER protein force field and an empirical model of solvation pertaining to the hydrophobic effect, a reasonable prediction for the A-chain of uteroglobin that has the right fold and α-carbon r.m.s.d of 7.4 Å. We have obtained a reasonable prediction for the DNA binding protein 1pou with a 6.3 Å α-carbon r.m.s.d, a structure that is our lowest energy value determined thus far, but which is still higher in energy than the crystal structure. We note that our approach is in no way limited by size and we hope to report on four additional α-helical target proteins ranging in size between 104 and 154 amino acids in the very near future.

A critical part of future work is the prediction of the more difficult β-sheet class, or arbitrary class, of proteins. This requires the extension of the use of restraints to include β-sheet and disulfide bond formation and regions where reverse turns occur and maybe even supersecondary structure motifs. We will use either our own developed network approach, or a variety of established structure prediction programs for predicting different aspects of structure such as protein class or disulfide bonds. This extension will also use the restraints to guide the global optimization algorithm, but now for other protein classes than α-helix.

Acknowledgements

Head-Gordon and Crivelli gratefully acknowledge support from U.S. Department of energy contract number DE-AC-03-76SF00098 and the national energy research supercomputer center (NERSC) for significant T3E computer resources. Byrd, Eskow and Schnabel gratefully acknowledge support from air force office of scientific research grant F49620-97-1-0164, army research office grant DAAG55-98-1-0176 and national science foundation grant CDA-9502956.

References

Assa-Munt, N., Mortishire-Smith, R.J., Aurora, R., Herr, W., Wright, P.E., 1993. Cell 73, 193.

Azmi, A., Byrd, R.H., Eskow, E., Schnabel, R., Crivelli, S., Philip, T.M., Head-Gordon, T., 1999. International conference on optimization in computational chemistry and molecular biology: local and global approaches. accepted for publication.

Bally, R., Delettre, J., 1984. J. Mol. Biol. 206, 153.

Byrd, R.H., Eskow, E., Oldenkamp, B., Schnabel, R.B., van der Hoek, A., 1995a. In: Proceedings of the Seventh SIAM Conference on Parallel Processing for Scientific Computing, SIAM, pp. 72–77.

Byrd, R., Eskow, E., van der Hoek, A., Schnabel, R., Shao, C.-S., Zou, Z., 1995b. In: Pardalos, P., Shalloway, D., Xue, G. (Eds.), Proceedings of the DIMACS Workshop on Global Minimization of Nonconvex Energy Functions: Molecular Conformation and Protein Folding, American Mathematical Society.

Byrd, R.H., Derby, T., Eskow, E., Oldenkamp, K.P.B., Schnabel, R.B., 1994. Large-Scale Optimization. In: Hager, W., Hearn, D., Pardalos, P. (Eds.), State Of The Art. Kluwer Academic Publishers, Dordrecht, pp. 71–84.

Cornell, W.D., Cieplak, P., Bayly, C.I., et al., 1995. J. Am. Chem. Soc. 117, 5179.

Crivelli, S., Gordon,T., Byrd, R.H.,Lecture Notes in Computer Science. In; Amestoy, P., Berger, P., Dayde, M., Duff, I., Fraysse, V., Giraud, L., Ruiz, D. (Eds.),Europar '99. pp. 578–585

Crivelli, S., Head-Gordon, T., 1999. J. Parallel Comput..submitted.

Eisenhaber, F., Persson, B., Argos, P., 1995. Crit. Rev. Biochem. Mol. Biol. 30, 1.

Friesner, R.A., Gunn, Jr., 1996. Ann. Rev. Biophys. Biomol. Struct. 25, 315.

Gay, D.M., Head-Gordon, T., Stillinger, F.H., Wright, M.H., 1992. Forefronts. Corn. Theory Center 8, 4.

Head-Gordon, T., Arrecis, J., Stillinger, F.H., 1991. Proc. Natl. Acad. Sci. USA 88, 11076.

Head-Gordon, T., Stillinger, F.H., Wright, M.H., Gay, D.M., 1992. Proc. Natl. Acad. Sci. USA 89, 11513.

Head-Gordon, T., Stillinger, F.H., 1993a. Biopolymers 33, 293.

Head-Gordon, T., Stillinger, F.H., 1993b. Phys. Rev. E. 48, 1502.

Head-Gordon, T., 1996. In: Merz, K.M., Le Grand, S.M. (Eds.), The Protein Folding Problem and Tertiary Structure Prediction. Birkhauser, Boston.

Head-Gordon, T., 1995. J. Am. Chem. Soc. 117, 501.

Head-Gordon, T., Sorenson, J.M., Pertsemlidis, A., Glaeser, R.M., 1997. Biophys. J. 73, 2106.

Holley, L.H., Karplus, M., 1991. Methods Enzym. 202, 204.

Hura, G., Sorenson, J.M., Glaeser, R.M., Head-Gordon, T., 1999. Perspectives in Drug Discovery. in press.

Kneller, D.G., Cohen, F.E., Langridge, R., 1990. J. Mol. Biol. 214, 171.

Liu, D.C., Nocedal, J., 1989. Math. Program. 45, 503.

Lovasz, L., Plummer, M.D., 1986. Matching Theory. Elsevier, Amsterdam.

McGregor, M.J., Flores, T.P., Sternberg, M.J.E., 1989. Protein Eng. 2, 521.

Muskal, S.M., Kim, S.-H., 1992. J. Mol. Biol. 225, 713.

Novotny, J., Bruccoleri, R., Karplus, M., 1984. J. Mol. Biol. 177, 787.

Pertsemlidis, A., Saxena, A.M., Soper, A.K., Head-Gordon, T., Glaeser, R.M., 1996. Proc. Natl. Acad. Sci. 93, 10769.

Pertsemlidis, A., Soper, A.K., Sorenson, J.M., Head-Gordon, T., 1999. Proc. Natl. Acad. Sci. 96, 481.

Pratt, L.R., Chandler, D., 1977. J. Chem. Phys. 67, 3683.

Pratt, L.R., Chandler, D., 1980. J. Chem. Phys. 73, 3434.

Qian, N., Sejnowski, T.J., 1988. J. Mol. Biol. 202, 865.

Rick, S.W., Berne, B.J., 1997. J. Phys. Chem. B. 101, 10488.

Rinnooy-Kan, A.H.G., Timmer, G., 1984. In: Boggs, P., Byrd, R., Schnabel, R.B. (Eds.), Numerical Optimization. SIAM, Philadelphia, pp. 245–262.

Rost, B., Sander, C., 1994. Proteins 19, 55.

Schiffer, C.A., Caldwell, J.W., Kollman, P.A., Stroud, R.M., 1993. Mol. Simul. 10, 121–149.

Sorenson, J.M., Head-Gordon, T., 1998. Fold. Des. 3, 523–534.

Sorenson, J.M., Hura, G., Pertsemlidis, A., Soper, A.K., Head-Gordon, T., 1999. J. Phys. Chem. B. 103, 5413.

Stillinger, F.H., Head-Gordon, T., Hirshfeld, C.L., 1993. Phys. Rev. E. 48, 1469.

Szebenyi, D.M.E., Moffat, K., 1986. J. Biol. Chem. 261, 8761.

Vasquez, M., Nemethy, G., Scheraga, H.A., 1994. Chem. Rev. 94, 2183.

Wesson, L., Eisenberg, D., 1992. Protein Science 1, 227–235.

Yu, R.C., Head-Gordon, T., 1995. Phys. Rev. E. 51, 3619.

Zichi, D.A., Rossky, P.J., 1985. J. Chem. Phys. 83, 797.